The Company of **Biologists**

RESEARCH ARTICLE

Identification and *in silico* modeling of enhancers reveals new features of the cardiac differentiation network

Denis Seyres[1,2], Yad Ghavi-Helm[3,*], Guillaume Junion[4,*], Ouarda Taghli-Lamallem[4], Céline Guichard[1,2], Laurence Röder[1,2], Charles Girardot[3], Eileen E. M. Furlong[3,‡] and Laurent Perrin[1,2,5,‡]

ABSTRACT

Developmental patterning and tissue formation are regulated through complex gene regulatory networks (GRNs) driven through the action of transcription factors (TFs) converging on enhancer elements. Here, as a point of entry to dissect the poorly defined GRN underlying cardiomyocyte differentiation, we apply an integrated approach to identify active enhancers and TFs involved in *Drosophila* heart development. The *Drosophila* heart consists of 104 cardiomyocytes, representing less than 0.5% of all cells in the embryo. By modifying BiTS-ChIP for rare cells, we examined H3K4me3 and H3K27ac chromatin landscapes to identify active promoters and enhancers specifically in cardiomyocytes. These *in vivo* data were complemented by a machine learning approach and extensive *in vivo* validation in transgenic embryos, which identified many new heart enhancers and their associated TF motifs. Our results implicate many new TFs in late stages of heart development, including Bagpipe, an *Nkx3.2* ortholog, which we show is essential for differentiated heart function.

KEY WORDS: Bagpipe, Nkx3.2, *Drosophila* heart, SVM modeling, Chromatin landscape, Differentiation network, Enhancer

INTRODUCTION

Understanding the molecular networks that regulate cell fate decisions and differentiation programs during embryogenesis is a key goal of developmental biology. Development is driven by very specific patterns of gene expression, which are largely regulated through the action of enhancer elements. Enhancers integrate information from multiple transcription factors (TFs) to give rise to a specific spatio-temporal expression pattern. As genes have many such enhancers, these elements function as regulatory nodes within large interconnected gene regulatory networks (GRNs) (Peter and Davidson, 2011). A comprehensive knowledge of enhancer activity at a genome-wide level is therefore an important first step in understanding how a GRN regulates a specific process or cell type.

The *Drosophila* heart provides an excellent model system with which to attain a complete systems level understanding of a tissue's lineage commitment and differentiation. It consists of a single

[1]Inserm UMR_S 1090, TAGC, Parc Scientifique de Luminy, Case 908, Cedex 9, Marseille 13288, France. [2]Aix-Marseille Université, TAGC, Parc Scientifique de Luminy, Marseille 13288, France. [3]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstr. 1, Heidelberg 69117, Germany. [4]Genetic Reproduction and Development Laboratory (GReD), INSERM U1103, CNRS UMR6293, Clermont-Ferrand 63000, France. [5]CNRS, Marseille 13009, France.
*These authors contributed equally to the work

‡Authors for correspondence (furlong@embl.de; laurent.perrin@univ-amu.fr)

Y.G., 0000-0003-2397-9576; C.G., 0000-0003-4301-3920; E.E. M.F., 0000-0002-9544-8339; L.P., 0000-0002-4139-8743

cardiac tube with two rows of cardiomyocytes made of 52 cell pairs, which form a lumen or cardiac cavity to pump the hemolymph around the body in an open circulatory system. These cardiac cells (CCs) are associated with non-muscle pericardial cells (PCs), which have common ontological origins. Despite divergent morphogenesis, the genetic network regulating heart development is highly conserved from flies to humans. For example, members of the same TF families and signaling pathways are essential for heart development [Tinman/Nkx2.5, H15Mid/Tbx, Gata, Hand, Dpp/BMP, Wg/Wnt (Reim and Frasch, 2010)] and are even similarly organized within the cardiac GRN (Olson, 2006).

Using two complementary approaches, recent genomic studies of cardiac specification in *Drosophila* have revealed a number of important features (Jin et al., 2013; Junion et al., 2012; Busser et al., 2015; Ahmad et al., 2014). First, genome-wide TF ChIP occupancy analyzed during stages of lineage commitment identified a large number of putative cardiac enhancers. These datasets indicate that there is little evidence for a specific TF motif architecture (often referred to as motif grammar) in cardiac enhancers (Jin et al., 2013; Junion et al., 2012). In fact, regions bound by five TFs essential for heart development not only lacked a consistent motif arrangement, but also often lacked the motif for some of the occupying TFs. Once a subset of TFs (three or four) is bound, they seem to provide a sufficient protein interface to recruit additional factors as a TF collective (Junion et al., 2012). Second, machine learning has been used for the genome-wide prediction of heart and pericardial enhancers (Ahmad et al., 2014) and enhancers active in subsets of CCs (Busser et al., 2015) based on a hand-curated training set of enhancers with characterized activity. Analysis of positive classifiers revealed novel transcription factor binding sites (TFBSs) and cognate TFs involved in the cardiac GRN (Ahmad et al., 2014) and sequence features that distinguished enhancers in different cell subpopulations (Busser et al., 2015). The number of new TFs that these studies have connected to heart development is remarkable (Busser et al., 2015): a total of 82 TFs, representing a tenfold excess compared with the number of characterized cell states. This highlights how much we still have to learn about the regulatory properties and function of enhancers within the cardiac GRN.

Previous studies have mainly focused on the early steps of cardiac development (cardiomyocyte specification). There is, however, much less information available about the cardiac GRN at later stages. In particular, to date only a few cardiac-specific enhancers and cognate TFs have been characterized during cardiac differentiation. ChIP-seq using antibodies directed against different chromatin modifications is a very effective method to identify cardiac differentiation enhancers genome-wide. We previously demonstrated, for example, that histone H3 acetylated on lysine27 (H3K27ac), H3 tri-methylation on lysine 79 (H3K79me3) and RNA polymerase II are highly predictive of enhancers in an active state (Bonn et al., 2012a,b). The identification

DEVELOPMENT

of enhancers based on chromatin modifications is also more comprehensive than enhancer discovery based on ChIP directed against specific TFs (Jin et al., 2013; Junion et al., 2012) or based on *in silico* modeling (Busser et al., 2015) as it does not require prior knowledge of the underlying TFs involved. However, at later stages of development during heart differentiation, CCs represent as little as 0.5% of the total cell population of the embryo, making tissue-specific experiments for enhancer identification a major challenge.

Here, we addressed these issues by modifying our previously published BiTS-ChIP method (Bonn et al., 2012a) to isolate tissue-specific chromatin specifically from this very rare cell population. We mapped H3K27ac- and H3K4me3-enriched regions to identify potential cardiomyocyte-specific active promoters and active enhancers. We complemented this *in vivo* approach with an *in silico* analysis of intergenic and intronic H3K27ac-enriched regions using a supervised machine learning approach to globally predict cardiac enhancers and identify their specific sequence features. Extensive *in vivo* validation revealed that many of these newly identified elements are active during cardiomyocyte differentiation, as predicted, implicating many new TFs in the GRN controlling cardiac differentiation. In particular, we uncovered a new role for Bagpipe (Bap), a TF known to function in visceral mesoderm specification but with no known role in late cardiogenesis. Taken together, this study provides a comprehensive insight into transcriptional regulation during cardiomyocyte differentiation, vastly extending the number of regulatory elements known to be active during these stages of heart development.

RESULTS
Tissue-specific ChIP on rare cell populations from developing embryos
BiTS-ChIP is an optimized protocol to obtain intact nuclei from fixed embryos, thereby preserving transcriptional and chromatin states during the dissociation and isolation procedure, combined with nuclei sorting using FACS, facilitating the isolation of stage- and tissue-specific chromatin at very high purity. The method was initially optimized to sort mesodermal nuclei, which represent ~20% of the developing embryo, and was used to probe chromatin states at active and inactive enhancers (Bonn et al., 2012a) and three-dimensional enhancer topology (Ghavi-helm et al., 2014). Here we improved BiTS-ChIP in two ways: first, to isolate nuclei from a very rare cell population, namely the 104 cardiomyocytes present per embryo; and second to perform ChIP on much smaller amounts of chromatin (Fig. 1A-C′). A transgenic *Drosophila* strain expressing nuclear GFP under the control of a cardiac-specific enhancer [TinC*>GFP (Hollfelder et al., 2014), Fig. 1B] was used for staged embryo collections at stages 13-14 (10-13 h of development). After embryo fixation, nuclei dissociation and fluorescent labeling, purification of this rare nuclei population was achieved by a two-step nuclear sorting procedure: a first high-speed enrichment sort that gave ~70% purity followed by a second high-accuracy slow sort that yielded ~98% purity (Fig. 1C,C′). Starting from 3 g of staged embryos, 7 million nuclei were obtained after the first sort and 3.5 million after the second sort. Chromatin was extracted and used for immunoprecipitation and sequencing (ChIP-seq) to analyze chromatin modifications at promoters (H3K4me3 and H3K27ac) and *cis*-regulatory elements (H3K27ac). This required optimization of our previous ChIP protocol to accommodate smaller amounts of chromatin (see Materials and Methods).

Two independent biological replicates (from embryo collections, FACS sorting, chromatin preparations and ChIP-seq) were analyzed, showing high concordance with each other with

Pearson correlation coefficients for H3K27ac and H3K4me3 reads of 0.98 and 0.88, respectively (see Materials and Methods). Fig. 1D shows H3K4me3 and H3K27ac signals obtained from cardiac chromatin at representative loci compared with signals obtained from mesoderm at an earlier developmental stage (Bonn et al., 2012a) and from stage-matched whole embryos (modENCODE data).

Cardiac BiTS-ChIP is highly specific
Promoters of genes known to have cardiac-specific expression at stages 13-16 showed a high enrichment for H3K4me3 and H3K27ac (Fig. 2A; *shg*, Berkeley Drosophila Genome Project, http://insitu. fruitfly.org/insitu_image_storage/img_dir_18/insitu18020.jpe), in contrast to those not expressed in the cardiac tube at this stage (Fig. 2B; *CG5171*, http://insitu.fruitfly.org/insitu_image_storage/ img_dir_7/insitu7206.jpe). For a more global analysis of the tissue specificity of our data, we used two large-scale datasets. First, a collection of 284 cardiac genes defined by Busser and colleagues (Busser et al., 2015) with confirmed spatiotemporal expression in the heart by *in situ* hybridization ('ISH gene set', Table S1). Promoters of genes expressed in the cardiac tube at stages 13-16 have high levels of both H3K27ac and H3K4me3 heart signals in the first nucleosomes in the body of the gene, as expected for genes with active transcription in this cell type (Fig. 2C, green). By contrast, promoters of genes that are not expressed in CCs, but do exhibit tissue-specific expression at the same stages of development in another tissue, have very low levels of heart-specific chromatin signatures (Fig. 2C, orange). Second, we performed RNA-seq on FACS-sorted CCs from staged transgenic embryos of a genotype identical to that used in the BiTS-ChIP experiments. Comparison with the stage-matched whole-embryo transcriptome (McKay and Lieb, 2013) allowed us to define genes with enriched cardiac expression with high accuracy ('RNA-seq gene set', see Materials and Methods, Fig. S1, Table S1). Both H3K4me3 and H3K27ac are also specifically enriched at promoters of genes expressed in CCs, as defined by RNA-seq (Fig. 2C′). Moreover, the level of chromatin signal is associated with the level of gene expression in cardiomyocytes (RPKM rank, Fig. 2D).

Both analyses demonstrate the tissue specificity of the data, obtaining chromatin signatures highly enriched for cardiomyocytes during late stages of heart development. However, given the very low amounts of chromatin (~500 ng) used for the ChIP, the recall is generally not as high as we would normally obtain. We applied stringent cutoffs to both the FACS gating and the procedure to call significantly enriched ChIP peaks. Peaks were defined using MACS2 (Zhang et al., 2008) and their consistency across replicates was filtered using irreproducible discovery rate (IDR) (Li et al., 2011) at a cutoff of 5% following the workflow recommended by modENCODE [see Materials and Methods (Landt et al., 2012) and Table S3]. We thus aimed to obtain data of high specificity to ensure a high true-positive rate, at the expense of recall, and therefore having a high false-negative rate. To examine the precision and recall we used the two sets of genes defined above, which are expressed in these cells at these stages of development (the ISH and RNA-seq gene sets), and analyzed the overlap of the peaks with their promoters (Table S4). As illustrated in Fig. 2E, both H3K27ac and H3K4me3 identify active heart promoters with high precision (76-89% depending on the gene set examined), in keeping with the analysis above. By contrast, their recovery (recall) varied from 21 to 28%, representing a tradeoff with the high specificity that we required and which might also illustrate subsaturating conditions due to the low amount of starting material used for ChIP-seq.
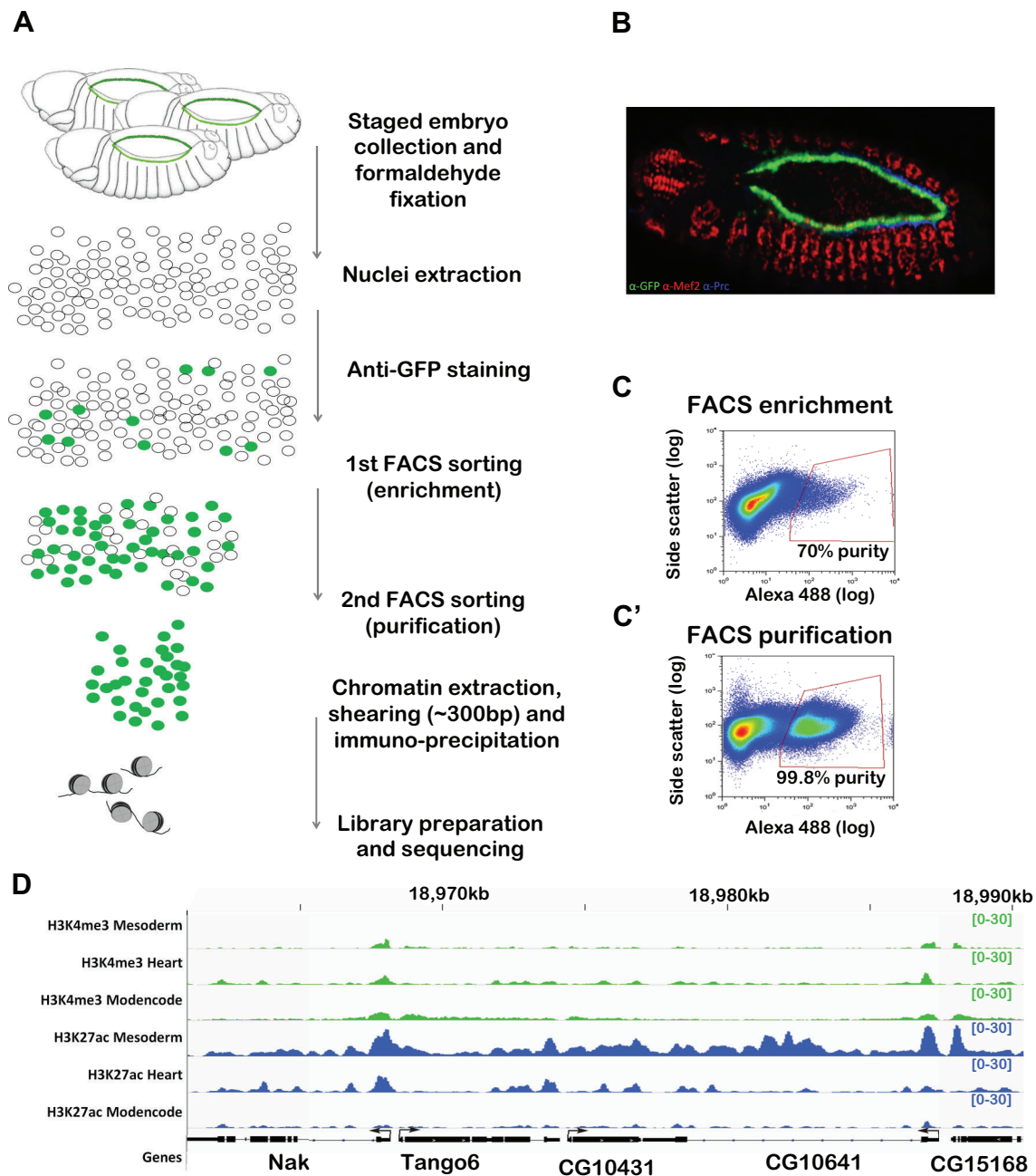
Fig. 1. Improved BiTS-ChIP allows cell type-specific chromatin profiling from scarce cell populations. (A) Method overview. Staged *Drosophila* embryos (10-13 h) expressing GFP specifically in cardiomyocyte nuclei are collected and fixed with formaldehyde. Nuclei are extracted and stained with a fluorescent antibody. FACS sorting to purity is achieved by combining an enrichment sort and a purification sort. Chromatin is extracted, sheared, immunoprecipitated and subjected to Illumina sequencing. (B) Stage 13 transgenic embryos expressing nuclear GFP under the control of the TinC* promoter. GFP (green) is specifically expressed in cardiomyocytes. Red, Mef2; blue, Pericardin (Prc). (C,C′) Representative FACS scatter plots of enrichment and purification sorts, relating side scatter (*y*-axis) to fluorescence intensity (*x*-axis). The red gates indicate the sorting events that were isolated. Nuclei sorted in C were resorted in C′ to achieve 99.8% purity. (D) Representative loci showing cardiac-specific H3K4me3 and H3K27ac profiles (heart) compared with stage-matched profiles obtained from whole embryos (modENCODE) and with those obtained using BITS-ChIP on mesodermal cells at 8-10 h (mesoderm) (Bonn et al., 2012a). *y*-axis shows library size-normalized reads per genomic coverage. Scale is indicated within square brackets.

Despite this low recovery, our data identified a large collection of promoters (defined as ±500 bp around annotated transcription start site, see the Materials and Methods) active in CCs (2165 or 2013 marked by H3K4me3 or H3K27ac, respectively, with 1056 with both marks; Table S4). The promoters harboring a significant peak for only one chromatin mark most likely reflect our subsaturating conditions.

Cardiac BiTS-ChIP identifies tissue-specific *cis*-regulatory modules (CRMs)

Given that the cardiomyocyte H3K27ac data can accurately identify active cardiac promoters, we next used the H3K27ac signal to identify putative active cardiac enhancer elements. We identified 322 high-confidence H3K27ac intergenic and intronic enriched regions (Table S4), with low H3K4me3 signal to eliminate potential
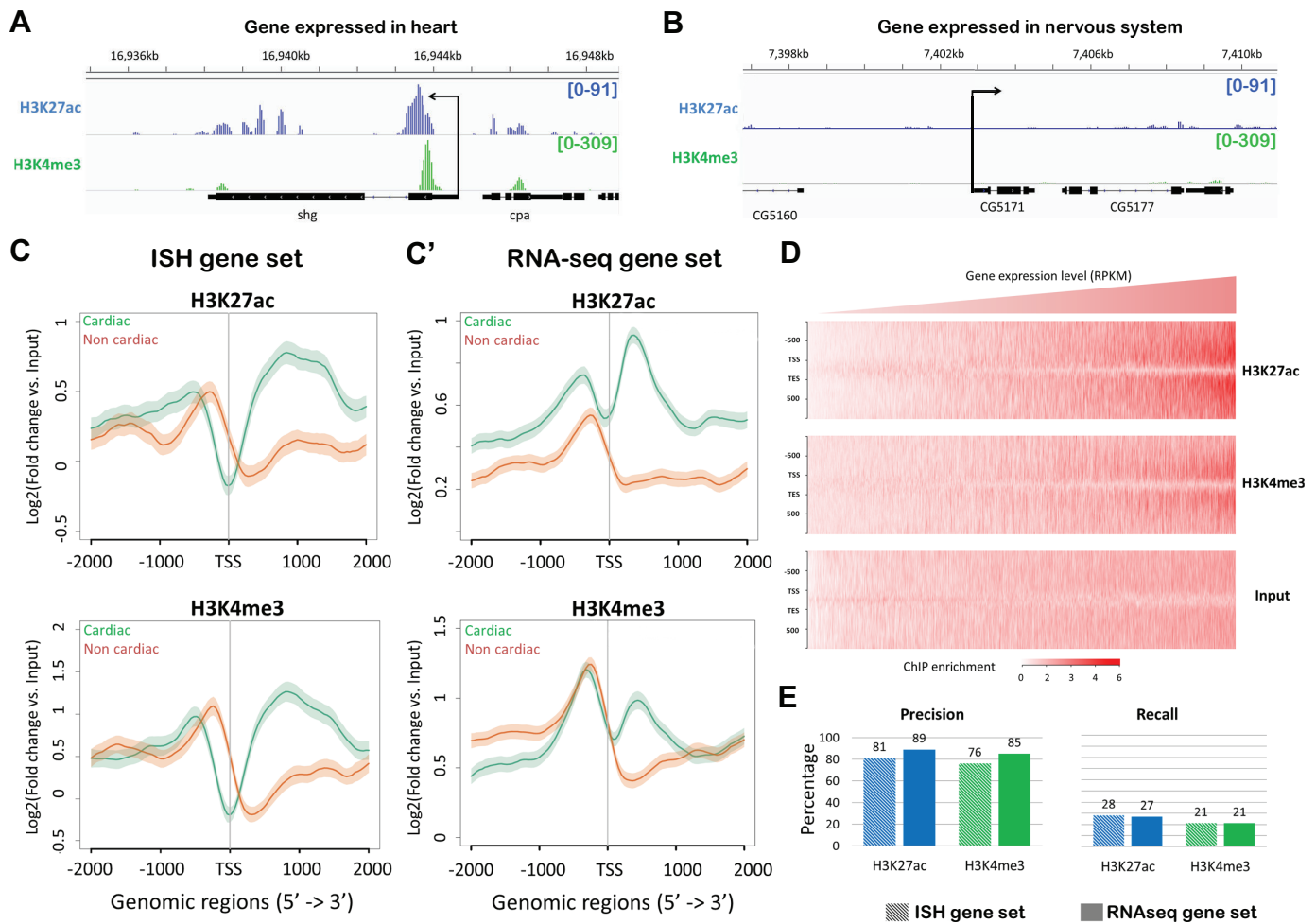
Fig. 2. Improved BiTS-ChIP has high tissue specificity. (A,B) BiTS-ChIP signal enrichment for H3K4me3 and H3K7ac marks for a cardiac (A, *shg*) and a non-cardiac (B, *CG5171*) expressed gene. Shown are background-subtracted reads per genomic coverage for H3K27ac (blue) and H3K4me3 (green). Arrow indicates transcription direction from the TSS. *y*-axis shows library size-normalized reads per genomic coverage. Scale is indicated within square brackets. (C,C′) Global assessment of tissue specificity. Background subtracted signal is shown for H3K27ac (top) and H3K4me3 (bottom) around the TSS of the cardiac ISH gene set (green, C) and RNA-seq gene set (this study, green, C′) compared with non-cardiac expressed gene sets (red). Both marks are specifically enriched at loci of cardiac expressed genes. (D) Chromatin mark levels on all *Drosophila* genes. Genes are ranked from more expressed (right) to less expressed (left) according to cardiac RNA-seq reads per kb per million mapped reads (RPKM). Intensity of ChIP signal is centered on gene bodies. (E) Specificity and recall of cardiac BiTS-ChIP. IDR defined peaks were associated to their nearest gene. Precision and recall for two independent sets of expressed genes (ISH genes set and RNA-seq genes set) for H3K4me3 or H3K27ac peaks.

unannotated promoters. Literature curation (Jin et al., 2013; Junion et al., 2012; Ahmad et al., 2014; Kvon et al., 2014; Pfeiffer et al., 2008) and the recent results of large-scale enhancer assays (Kvon et al., 2014) allowed us to establish a collection of 80 *in vivo* characterized enhancers driving expression in CCs and a set of 105 enhancers active in another tissue (non-cardiac) at the same developmental stages 13-16 (Table S2). As shown in Fig. 3A-C, the BiTS-ChIP signal for H3K27ac in cardiomyocytes is enriched at cardiac enhancers but not at enhancers active in non-cardiac tissues, demonstrating the specificity of the data to identify heart enhancers. More globally, intergenic and intronic H3K27ac peaks identified cardiac enhancers with high precision (87.5%, Fig. 3D), again demonstrating the specificity of the data. The recovery was relatively low (18.9%, Fig. 3D), indicating reduced recall as expected given the stringent cutoffs applied. Bonn et al. (2012b) established that not all active enhancers are marked by H3K27ac in the developing mesoderm, which might also be the case in differentiating CCs, which would additionally impact the estimated recovery. Importantly, the genes associated with

intergenic/intronic H3K27ac peaks are enriched in specific functional annotations that are highly relevant to cardiac differentiation and morphogenesis, including cell and tissue morphogenesis and heart development (Fig. 3E).

As the main aim of this study was to identify new heart enhancers active during late stages of embryogenesis, we examined the ability of new genomic regions encompassing intergenic or intronic H3K27ac peaks to drive cardiac-specific expression at stages 13-16 *in vivo* in transgenic embryos. We tested 15 regions encompassing IDR defined peaks in the vicinity of *in situ* defined heart-expressed genes (Table S3). The cardiac activity of three of these regions has also recently been analyzed by Busser et al. (2015). Twelve out of the 15 (80%) regions tested function as developmental enhancers at these stages of embryogenesis; 58% of these (7/12) regulate expression in all or a subset of differentiating cardiomyocytes (Fig. 3F, Fig. S2A), while the remaining five regions regulate expression in other tissues, including the nervous system, amnioserosa, and subsets of ectodermal cells (Fig. S2A). These scores compare well to those published by Busser et al. (2015) at

DEVELOPMENT
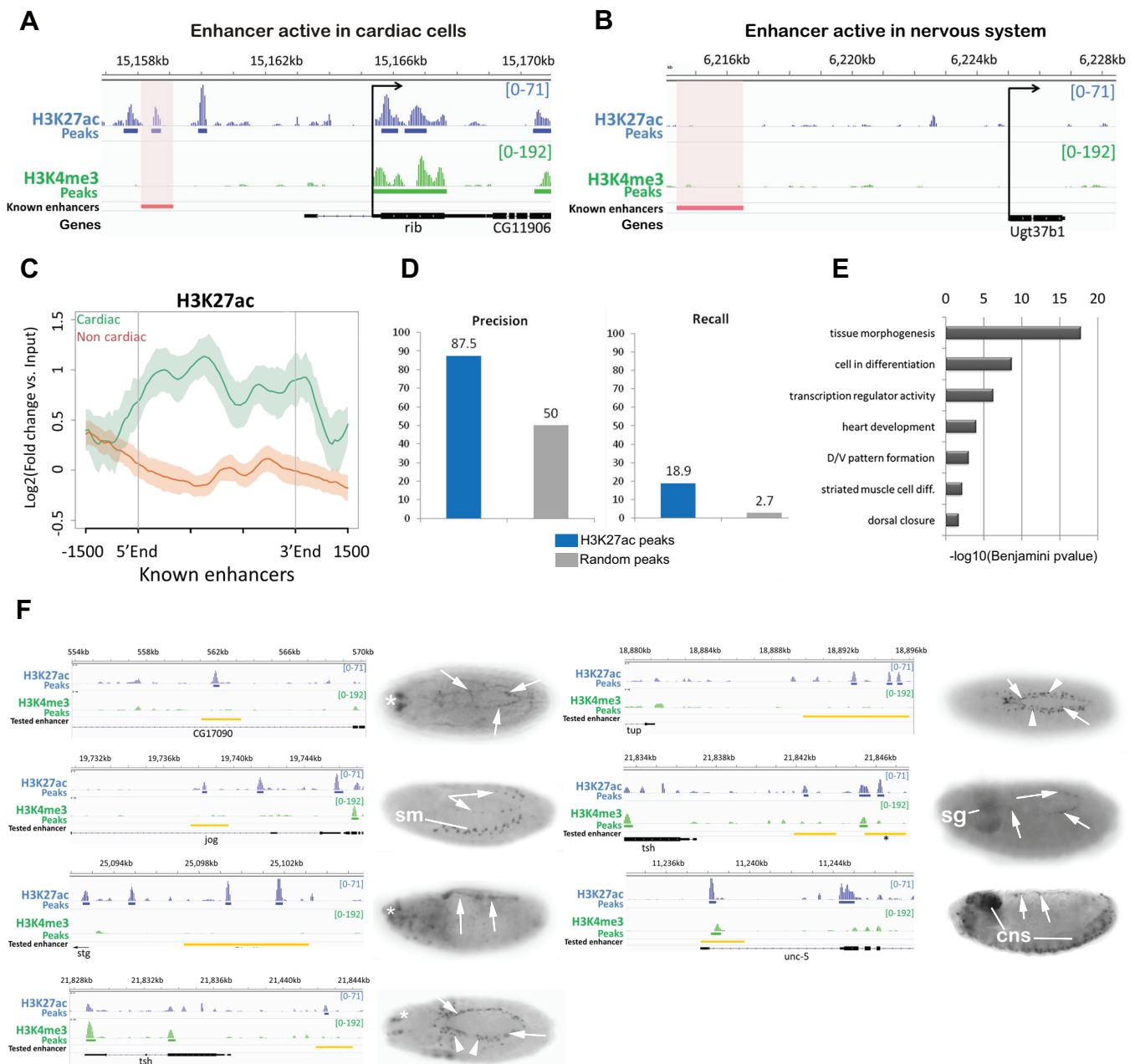
Fig. 3. H3K27ac cardiac BiTS-ChIP accurately identifies tissue-specific CRMs. (A,B) Cardiac BiTS-ChIP signal enrichment for validated CRMs driving tissue-specific expression in the cardiac tube (A) or excluding the cardiac tube (B) at 10-13 h of embryonic development. Background-subtracted signal and IDR called peaks are shown, together with genomic regions tested in reporter gene assays (red box). (C) Global assessment of tissue specificity of H3K27ac signal on CRMs. Cardiac CRMs (green) display enriched H3K27ac signal compared with stage-matched non-cardiac CRMs (orange). (D) Precision and recall for *in vivo* validated active cardiac enhancers identified with H3K27ac cardiac BiTS-ChIP at 10-13 h (blue). Both precision and recall fall when random peaks are used (negative control; gray) (see Materials and Methods). (E) Functional annotations of nearest genes of intergenic/intronic H3K27ac peaks. GO (biological process) term enrichment (*P*-value cutoff: 0.05) is shown. (F) H3K27ac-enriched regions function as enhancers *in vivo*. (Left) Genomic regions encompassing intergenic/intronic H3K27ac peaks. Background-subtracted signal and IDR called peaks are shown, together with genomic regions tested in reporter gene assays (yellow). *y*-axis shows library size-normalized reads per genomic coverage. (Right) Stage 13-16 embryos showing the expression pattern of the *Gal4* reporter gene driven by the genomic regions tested detected by *in situ* hybridization. Anterior is left and views are dorsal or sagittal-dorsal. Expression in cardiomyocytes is indicated by arrows, in pericardial cells by arrowheads. Other tissue-specific expressions are indicated: cns, central nervous system; sg, salivary glands; sm, somatic muscles; asterisks, tissue not determined. See Fig. S2A for additional data.

early stages of cardiac development, who reported that 95% of tested elements function as enhancers, with 41.3% (19/46) being active in the heart. Taken together, this analysis revealed that cell type-specific information on non-coding regions with high H3K27ac (and low H3K4me3) is a good indicator of cell type-specific enhancers, with almost 60% of active enhancers having the

expected cell type-specific expression. Although much higher than random, this is, interestingly, still much lower than the precision obtained when looking at a large collection of known enhancers (87.5%, Fig. 3D), suggesting that perhaps the known enhancers have a broader activity (all cardioblasts) or a higher level of H3K27ac signal, which we most likely underestimate here. Of note,

our approach might underestimate the proportion of active enhancers during late stages of heart development. Indeed, some of the tested regions might only work with their own proximal promoter or might require a more extended region to function than that cloned here.

Using genomic features and machine learning to predict cardiac enhancers genome-wide

Given our relatively low recovery of known cardiac enhancers (and genes), owing to the stringent criteria applied, we are clearly underestimating the genome-wide content of cardiac-specific enhancers (Fig. 3D). The subsaturating nature of our ChIP-seq dataset will therefore lead to a high proportion of false negatives. To identify more cardiac enhancers, we used a complementary approach: a support vector machine (SVM)-based supervised classification to model cardiac enhancers. We recently demonstrated that the LedPred R-package (Seyres et al., 2015), which is designed for supervised classification based on SVM, excels at the prediction of regulatory sequences. Using LedPred, the SVM was designed to identify combinations of TFBSs that optimally separate H3K27ac-containing putative cardiac enhancers

from control regions. We selected H3K27ac peaks in the vicinity (5 kb upstream and introns) of genes known to be expressed in cardiomyocytes. These defined a positive training set of 23 H3K27ac-positive intergenic/intronic regions (Table S6). As a negative training set, a pool of 460 enhancers not active in the heart but still driving stage-matched tissue-specific expression was used (see Materials and Methods, Table S6). A radial-kernel SVM classifier was used for enhancer classification, based on matching of position-weighted matrices (PWMs) characterizing TFBSs from the vertebrate and insect JASPAR collection (Mathelier et al., 2014). We assessed the classifier by its ability to predict the correct class by tenfold cross-validation using precision and recall curves. As a negative control, we trained four different SVM using sets of 23 regions randomly chosen among the 460 enhancers of the negative set and compared their average precision/recall curves with that of the positive training set (Fig. 4A). The SVM trained on cardiac H3K27ac-positive regions outperforms the SVMs trained on negative regions.

The trained SVM model was then used for genome-wide predictions of regulatory elements in order to identify genomic regions similar to the defined model. We predicted enhancers
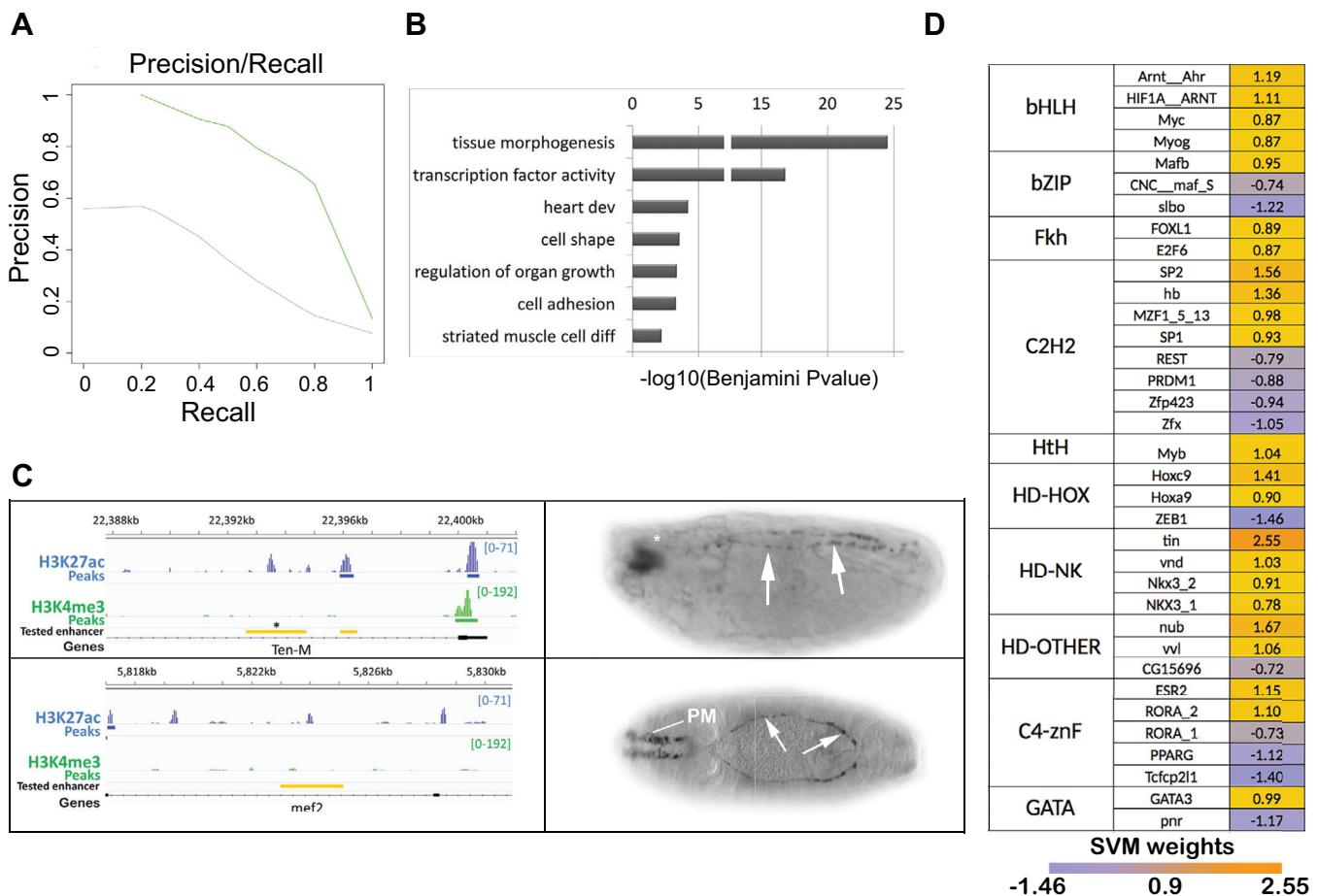
Fig. 4. H3K27ac peak classifier performs with high specificity and recall. (A) Precision/recall curve for radial SVM modeling of H3K27ac peaks in intergenic/intronic regions (green) and for SVMs trained on negative regions (gray; mean of four independent SVM each trained on 23 negative regions). (B) Enrichment in functional annotations of genes associated with cardiac CRMs predicted with the classifier ordered by decreasing Benjamini-corrected *P*-value. (C) *In vivo* validation of predicted cardiac CRMs. (Left) Genomic regions encompassing predicted cardiac CRMs. Background-subtracted signal of cardiac BiTS-ChIP is shown. *y*-axis shows library size-normalized reads per genomic coverage. Scale is indicated within square brackets. Yellow boxes indicate the genomic regions tested in the *in vivo* reporter gene assay. (Right) Stage 13-16 embryos showing reporter expression patterns detected by *in situ* hybridization. Anterior is left. Tested genomic regions drive transcription in the cardiomyocytes at the predicted stages of development (arrows). Other tissue-specific expressions are indicated: pm, pharyngeal muscle; asterisk, tissue not determined. (D) PWMs with the most extreme SVM weights are represented and classified by TF class. More SVM weight means that the PWM is more discriminative for the model.

DEVELOPMENT

genome-wide by splitting the entire non-coding *Drosophila* genome (dm3) into 1500 bp regions with 750 bp overlap and scoring the regions with the trained SVM model. We chose an estimated precision of 0.95 and a recall of 0.35 as a cutoff (from Fig. 4A, precision/recall curve, corresponding to an SVM score above 0.9) to select regions. This represents 2364 regions (Table S7).

Four lines of evidence suggest that these predicted regions are enriched in new cardiac enhancers, supporting the accuracy of the model. First, a high number (52/316) of intergenic/intronic regions containing H3K27ac peaks that were not included in the positive training set do overlap with a predicted region ($P=0.0001$; see supplementary Materials and Methods). Second, 17 SVM predicted regions correspond to characterized cardiac enhancers (i.e. are among of the 80 enhancers described above; $P<0.0002$, Table S4). Third, the model retrieves genomic regions that are located close to 69 genes known to be expressed in differentiating cardiomyocytes (Fisher's test, $P=5.6×10^{-17}$), which are part of the 284 ISH gene set. Fourth, genes in the proximity of predicted regions (1404 genes) are significantly enriched in biological functions [gene ontology (GO) terms] expected for cardiac differentiation: heart development ($P=0.0001$), striated muscle cell differentiation ($P=0.001$), cell shape ($P=0.0001$), cell adhesion ($P=0.001$), TF activity ($P=1×10^{-16}$) and tissue morphogenesis ($P=1×10^{-24}$) (Fig. 4B). Importantly, when we repeated this GO analysis using our random SVMs, general terms associated with development were reported (tissue morphogenesis and TF activity) but none of the specific heart terms. These results confirm the ability of LedPred to predict regulatory regions and indicate that the genome-wide predictions of enhancers from the model are highly enriched for cardiac-specific activity.

To challenge the SVM classifier, we selected four predicted regions with a high SVM score (over 0.9) but with no experimentally defined H3K27ac peaks, and tested their activity *in vivo* using transgenic reporter assays. Two of the four regions tested were sufficient to drive cardiac-specific expression at stages 13-15 (Fig. 4C), confirming the SVM prediction. These two regions might well be marked by H3K27ac but were missed in our analysis due to the subsaturating nature of our ChIP data, highlighting the usefulness of the SVM modeling as a complementary approach to ChIP-seq. Interestingly, the other two tested regions also function as enhancers at similar developmental stages, but drive reporter gene expression in CNS and visceral mesoderm (Fig. S2B).

TFBS features associated with cardiac enhancers
The enrichment of the genome-wide predictions for cardiac enhancer activity suggests that the sequence features learned by the SVM are robust enough to learn TF motif signatures in cardiac enhancers and therefore to potentially identify new TFs involved in cardiac differentiation. To investigate this, we examined the sequence features of the training set used for the model. When using SVM, features specifically associated with the positive set were given a positive weight, whereas those associated with the control set were given a negative weight (Table S8). Of interest, features that were given a positive weight in the model (Fig. 4D, Table S8) include PWMs for Tinman (Tin) and GATA, two TFs known to be essential for heart development. Other features include potential PWMs for TFs that do not have a characterized role in the cardiac GRN. These include a motif for the vertebrate TF Nkx3.2, an ortholog of *Drosophila* Bagpipe (Bap).

The good ranking of the Nkx3.2 PWM suggests that Bap is involved in late cardiac differentiation. During earlier developmental stages, *bap* is expressed in the prospective visceral mesoderm (VM), where it is required for VM specification, which in turn represses cardiac mesoderm (Azpiazu and Frasch, 1993). The same study established cardiac expression of *bap* at later stages (Azpiazu and Frasch, 1993), although no functional role for *bap* in heart development has been reported to date. To further probe the possible involvement of Bap motifs in cardiac enhancers, we analyzed the potential motifs specifically harbored by cardiac enhancers by performing motif discovery on 29 *in vivo* validated cardiac enhancers overlapping H3K27ac intergenic/intronic peaks (Table S9). Potential TFBSs for Tin, Bap, Seven up (Svp) and Ladybird early (Lbe) are enriched in these sequences. These motifs were also retrieved when motif discovery was performed in orthologous sequences on eight other *Drosophila* species, indicating that they are conserved and therefore might constitute functional motifs, but were not retrieved from random sequences (Fig. S3A). The recovery of Bap motifs was not due to similarity between Tin and Bap motifs (two TFs of the Nkx family) since mapping both motifs in the validated cardiac enhancers clearly showed that they do not overlap (Fig. S3B).

Bap is a novel member of the cardiogenic network
To determine if Bap is essential for heart development, we analyzed *bap* expression in the developing cardiac tube at both the RNA and protein levels. Importantly, as previously reported, *bap* is not expressed during cardiomyocyte specification (stage 10/11, not shown), and both RNA and protein were detected in cardiomyocytes from stage 13 onward (Fig. 5, Fig. S4). First detected in two out of four Tin-expressing CCs per hemisegment at stage 13 (Fig. 5A-C), *bap* expression is activated in all Tin+ CCs at stage 14 (Fig. 5D-F) and this expression is maintained to adulthood (Fig. 5G-I).

To examine the function of *bap* in cardiogenesis we knocked down *bap* specifically in cardiac muscle and analyzed both structural and functional consequences. *bap* loss-of-function greatly impacted myofibrillar organization and content in the mature heart (Fig. 5J,K). In particular, *bap* knockdown hearts displayed a much less compact arrangement of myofibrils than controls. To study heart function, we dissected flies in artificial hemolymph to record cardiac contractions (Ocorr et al., 2009). The *bap*-deficient hearts exhibited a significantly increased heart rate and also displayed increased variation in heart periodicity, quantified using the heart period standard deviation [arrhythmia index (AI)] (Fig. 5L). We further probed *bap* involvement in cardiac function by measuring fractional shortening, which is the proportional decrease in heart wall diameter during contraction and provides an indication of cardiac output. Fractional shortening in control flies was 40-45%, whereas it was reduced to 25-30% in *bap* knockdown hearts (Fig. 5L), further illustrating *bap* requirement for wild-type cardiac function. This functional analysis therefore revealed a central role for the evolutionary conserved TF Bap in cardiomyocytes differentiation and function.

DISCUSSION
In this study, tissue-specific enhancers (and promoters) were identified based on the presence of H3K27ac and absence of H3K4me3. Most previous studies in *Drosophila* have used the binding pattern of known cardiogenic TFs to identify cardiac-specific enhancers genome-wide (Jin et al., 2013; Junion et al., 2012). However, this approach can only be applied to well-
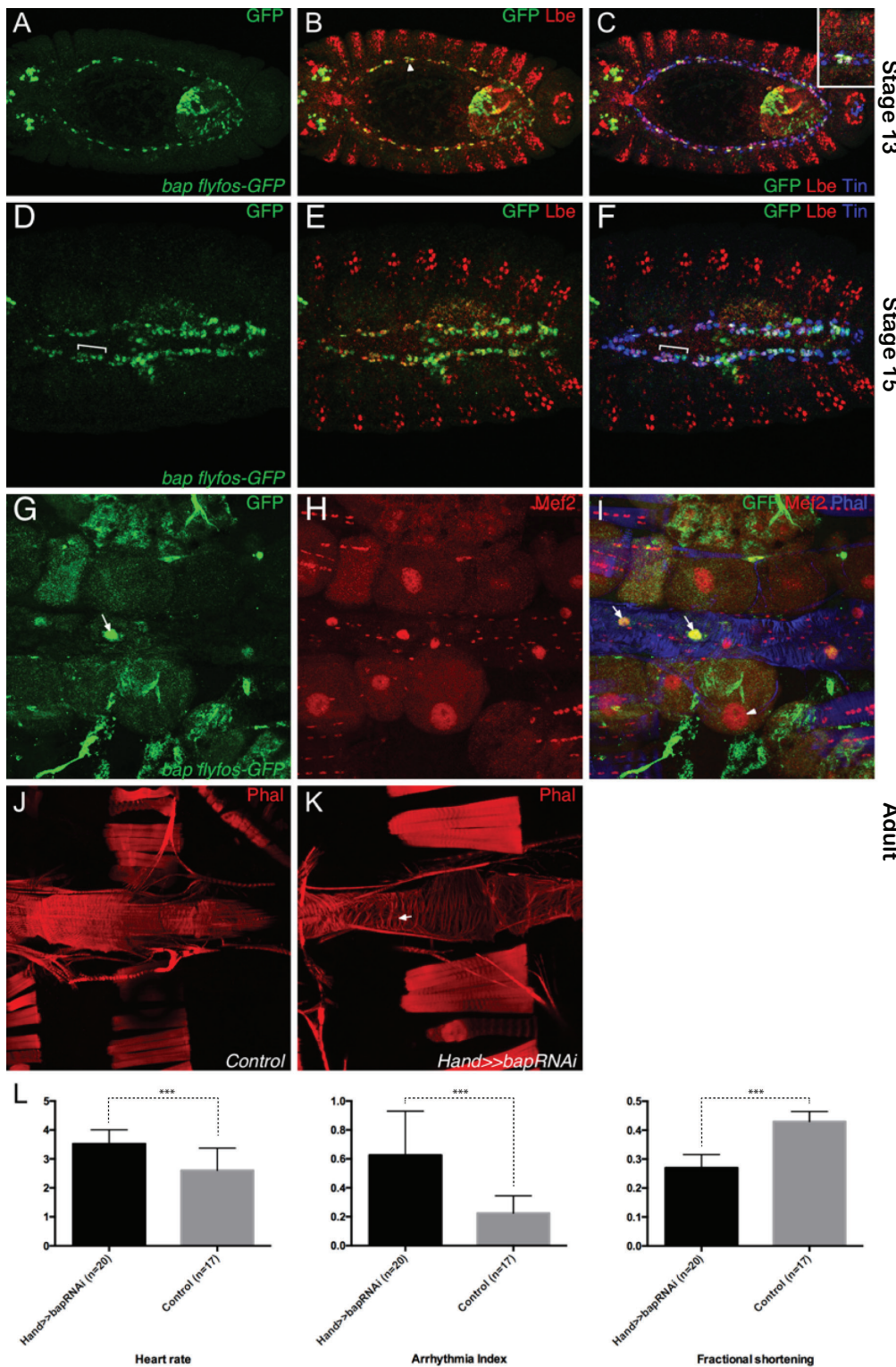
Fig. 5. Bap expression and function in embryonic and adult cardiomyocytes. (A-C) GFP-tagged FlyFos transgenic embryo showing Bap protein expression beginning in two out of four Tin[+] cardioblasts and two pericardial cells (PCs) expressing Lbe (arrowhead) at late stage 13. Inset (C) shows Lbe expression at higher magnification. (D-F) Bap-GFP became localized in four Tin[+] cardioblasts and two Lbe[+] PCs at embryonic stage 15 (bracket). This expression was maintained until the end of embryogenesis. (G-I) Bap-GFP expression in adult (1-week-old) dissected heart tube stained with phalloidin (blue) showing expression in some Mef2-expressing cardiomyocytes (arrows). Note that PC expression is lost in the adult fly (arrowhead). (J,K) Structural organization of cardiomyocytes in *bap* knockdown adult fly hearts labeled with phalloidin shows strong cardiomyocyte compaction defects (arrow) compared with control heart. (L) Functional monitoring of cardiac function in Bap-deficient adult heart shows an increase in heart rate and variability in heart period (arrhythmia index, AI) and a strong decrease in fractional shortening compared with control flies of the same age (1 week). For statistical analysis, unpaired Student's *t*-tests with Welch's correction were performed for each parameter, showing significant differences: heart rate, ***P=0.0002; AI, ***P<0.0001; fractional shortening, ***P<0.0001. Error bars indicate s.d. These data show a requirement for Bap function during adult heart differentiation.

characterized TFs with a known role in the tissue of interest, and for which an antibody is available. ChIP against TFs will also only identify a subset of the enhancers involved, unless a clear master regulator of that tissue's development is known. By contrast, H3K27ac is thought to unravel functionally active enhancers (Bonn et al., 2012b) without any prior knowledge of the TFs involved.

In this study, we show that presence of the H3K27ac mark is not an absolute predictor of tissue-specific activity, as even in purified population of cardiac nuclei only ~60% of identified active enhancers drive expression in this cell type. Although the biological reason for this is currently unclear (perhaps perdurance of this mark from an earlier developmental stage), it indicates that there is not a direct relationship between the presence of H3K27ac on an enhancer and enhancer activity, in agreement with our previous observations (Bonn et al., 2012b). In this study, the use of H3K27ac was also motivated by the fact that ChIP directed against

chromatin marks generally requires much less chromatin than ChIP against TFs, which was limiting in our case; indeed, the recovery of ChIP genomic DNA is often at least two orders of magnitude higher for chromatin marks than with TFs, allowing the quantity of chromatin required to be significantly downscaled. Of note, an alternative method, ATAC-seq (Buenrostro et al., 2013), which identifies nucleosome-free regions that are accessible to bacterial (Tn5) transposase, has recently been shown to be well suited for limited amounts of chromatin (Davie et al., 2015). It could therefore be used as a complementary approach to H3K27ac ChIP-seq on BiTS chromatin preparations. BiTS-ChIP on H3K27ac provides important information on the activity state of enhancers, which is not gained with ATAC-seq. However, ATAC-seq on BiTS sorted nuclei would improve the sensitivity of enhancer identification, as it would identify all elements regardless of whether they are active or not or harbor H3K27ac.

In addition, although the kernel of TFs that drive cardiac specification is well characterized and hence justifies the choice of TF ChIP to analyze the regulatory landscape of early cardiogenesis, much less was known regarding the regulators of cardiac differentiation. In this respect, the use of histone marks allowed an unbiased data-driven approach that enabled us to identify new players and features of the cardiac GRN.

Isolating tissue-specific chromatin was also key to acquiring well-defined spatiotemporal datasets. Here we chose to improve the BITS-ChIP-seq protocol, which is based on FACS sorting of tissue-specific nuclei (marked by either a transgene or antibody for a cell type-specific TF) from formaldehyde cross-linked embryos (Bonn et al., 2012a). Alternative methods were recently developed in *Drosophila* for tissue-specific nuclear isolation. In particular, the INTACT method is based on affinity purification of biotin-tagged unfixed nuclei (Steiner et al., 2012). When precise timing is a prerequisite, formaldehyde cross-linking is an important step in order to 'freeze' transcriptional states. The modified sorting strategy that we used (based on nuclear sorting using an initial enrichment sort) enabled accelerated purification of chromatin from a cell population of very low abundance. Starting from such a rare cell population (104 cardiomyocytes, among tens of thousands of cells in the whole embryo at late developmental stages), the amount of purified chromatin was always going to be limiting. As a consequence, we tuned the peak calling in order to achieve high specificity (low false-positive rates) but at the expense of recall, and therefore expect a great number of false negatives in our dataset. This nevertheless allowed us to identify tissue-specific active promoters and enhancers with very high precision (89% and 87.5%, respectively).

To enlarge the set of potential cardiac enhancers active at this developmental stage we used our genomic datasets together with supervised classification (SVM) to predict putative cardiac enhancers genome-wide. Importantly, modeling H3K27ac-marked regions was very effective and the model behaved as a good predictor of cardiac CRMs, as illustrated by *in silico* and *in vivo* validations. We also used the SVM classifier to uncover motif information within sets of sequences. For this purpose, we analyzed the positively weighted PWMs of the model trained on H3K27ac-bound regions. These included well-characterized motifs for cardiogenic TFs, including Tin and GATA. In addition, we identified the Nkx3.2 PWM, suggesting that the fly homolog Bap is involved in late cardiac differentiation. We indeed show that Bap is expressed in cardiomyocytes during their differentiation and in the mature heart and provide evidence that its function is required for proper morphogenesis and function of the heart. Therefore, while *bap* is specifically required for VM specification at earlier

embryonic stages (Azpiazu and Frasch, 1993), it is recruited to the cardiac GRN at later stages.

Our identification and modeling of cardiac H3K27ac-positive regions enables an accurate global prediction of cardiac enhancers and highlights new features of the cardiac GRN at late developmental stages. Our approach therefore constitutes a useful complementary approach to the previously reported modeling of heart enhancers, either in *Drosophila* (Busser et al., 2015) or humans (Narlikar et al., 2008), which started from characterized (*in vivo* validated) heart enhancers. Our success rate compares very well to that of a previous study that predicted heart enhancers active during early stages of development: 80% of our tested elements function as active developmental enhancers, with 58% being active in the heart; while 95% of the Busser et al. (2015) elements function as active elements, with 41.3% being active in CCs. The vertebrate study by Narlikar et al. (2008) recovered 62% of tested elements that are active in the heart. Overall, the performances are therefore comparable. However, each approach has its own goals and advantages. The approach taken by Narlikar et al. (2008) used characterized enhancers and an iterative modelisation step to extract a homogeneous set of heart enhancers for training, to predict heart enhancers with high accuracy. The study by Busser et al. (2015) also made use of *in vivo* validated cardiac enhancers, in this case those active in different CC types, allowing the identification of transcriptional regulatory signatures of individual CC states in *Drosophila*. Our approach was data-driven, without prior knowledge of validated cardiac enhancers and regulatory TFs, making it applicable to almost all systems. In addition, our approach should be able to identify a diverse set of heart enhancers, regulating expression in diverse CC subtypes, and not just those with homogenous activity.

Conclusions

Comprehensive identification of *cis*-regulatory sequences active in specific tissues during embryonic development is a major challenge. It is however a central issue if one wants to analyze GRNs that drive tissue or organ development. Here we used an approach that combines *in vivo* identification of tissue-specific enhancers and supervised classification to model *cis*-regulatory information involved in cardiac differentiation in *Drosophila*.

We analyzed cardiomyocyte-specific H3K4me3 and H3K27ac chromatin landscapes from developing embryos and identified thousands of active cardiac promoters and putative enhancers. Modeling, using machine learning, revealed new features of cardiac enhancers, and we established that the *Nkx3.2* ortholog *bap* is required for cardiac differentiation during late embryogenesis, in addition to its well-established role as a regulator of VM specification at earlier developmental stages.

These findings illustrate the importance of analyzing GRNs at different developmental stages by acquiring precise tissue- and stage-specific datasets. A number of other features that are outlined by the model certainly deserve future investigation. Our results therefore open new avenues in the analysis of the cardiac GRN.

MATERIALS AND METHODS
Cardiac BiTS-ChIP-seq
Batch isolation of cardiac nuclei and immunoprecipitation of chromatin were performed as described (Bonn et al., 2012a) with following modifications: FACS sorting included an enrichment sort before the purification sort and ChIP procedures were downscaled to accommodate as little as 500 ng input material. Two biological replicates at 10-13 h after egg laying were generated for each condition. Library preparation and sequencing were performed using NEBNext Kit and Illumina sequencing.

Peaks were defined using MACS2 and selected using IDR following the procedure described by Landt et al. (2012). For details of the embryo collection, sorting of nuclei, ChIP procedures and ChIP-seq data acquisition and analysis, see the supplementary Materials and Methods.

Cardiac RNA-seq

Cardiac cells were purified by FACS (Defaye and Perrin, 2014). mRNA was linearly amplified and sequenced on an Illumina HiSeq 2000. For details of cardiac RNA-seq data acquisition and analysis, see the supplementary Materials and Methods.

SVM and motif analysis

Supervised classification of H3K27ac peaks and cardiac enhancers was performed using LedPred (Seyres et al., 2015). DNA motif analysis was performed using the RSAT suite (Turatsinze et al., 2008). For details, see the supplementary Materials and Methods.

Transgenic reporter assays

CRM activity was assayed using transgenic reporter assays as previously reported (Junion et al., 2012). Transgenic lines were tested by *in situ* hybridization using probes against reporter genes. For details, see the supplementary Materials and Methods.

Heart assays

For physiological analysis, fly hearts were surgically exposed and cardiac contractions were analyzed according (Fink et al., 2009). For details, see the supplementary Materials and Methods.

Author contributions

D.S., Y.G.-H., G.J., O.T.-L., C. Guichard, L.R. and L.P. performed experiments. D.S., G.J., C. Girardot, E.E.M.F. and L.P. obtained and analyzed data. D.S., Y.G.-H., G.J., C. Girardot, E.E.M.F. and L.P. wrote the manuscript. E.E.M.F. and L.P. secured funding.

Data availability

Raw sequences and mapping to the *D. melanogaster* reference genome dm3 (BAM files) of heart-specific BiTS-ChIP-Seq data (for H3K4me3 and H3K27ac) and input-Seq data generated in this study are accessible at the Sequence Read Archive (SRA) under study accession number ERP015146. RNA-seq data are available under SRA accession number ERP015942. ChIP signals (i.e. the log2 ratio) for genome browser visualisation are also available as bigwig files from the Furlong lab (http://furlonglab.embl.de/).

Supplementary information

Supplementary information available online at
http://dev.biologists.org/lookup/doi/10.1242/dev.140822.supplemental

References

Ahmad, S. M., Busser, B. W., Huang, D., Cozart, E. J., Michaud, S., Zhu, X., Jeffries, N., Aboukhalil, A., Bulyk, M. L., Ovcharenko, I. et al. (2014). Machine learning classification of cell-specific cardiac enhancers uncovers developmental subnetworks regulating progenitor cell division and cell fate specification. *Development* **141**, 878-888.

Azpiazu, N. and Frasch, M. (1993). Tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila. *Genes Dev.* **7**, 1325-1340.

Bonn, S., Zinzen, R. P., Perez-Gonzalez, A., Riddell, A., Gavin, A.-C. and Furlong, E. E. M. (2012a). Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. *Nat. Protoc.* **7**, 978-994.

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A. and Furlong, E. E. M. (2012b). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148-156.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. A. and Greenlea, W. J., (2013). Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat. Methods* **29**, 1213-1218.

Busser, B. W., Haimovich, J., Huang, D., Ovcharenko, I. and Michelson, A. M. (2015). Enhancer modeling uncovers transcriptional signatures of individual cardiac cell states in Drosophila. *Nucleic Acids Res.* **43**, 1726-1739.

Davie, K., Jacobs, J., Atkins, M., Potier, D., Christiaens, V., Halder, G. and Aerts, S. (2015). Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet.* **11**, e1004994.

Defaye, A. and Perrin, L. (2014). Tissue specific RNA isolation in Drosophila embryos: a strategy to analyze context dependent transcriptome landscapes using FACS. *Methods Mol. Biol.* **1196**, 183-195.

Fink, M., Callol-Massot, C., Chu, A., Ruiz-Lozano, P., Belmonte, J., Giles, W., Bodmer, R. and Ocorr, K. (2009). A new method for detection and quantification of heartbeat parameters in Drosophila, zebrafish, and embryonic mouse hearts. *BioTechniques* **46**, 101-113.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W. and Furlong, E. E. (2014). Enhancer loops appear stable during development and are associatedwith paused polymerase. *Nature* **512**, 96-100.

Hollfelder, D. Frasch, M. and Reim, I. (2014). Distinct functions of the laminin β LN domain and collagen IV during cardiac extracellular matrix formation and stabilization of alary muscle attachments revealed by EMS mutagenesis in Drosophila. *BMC Dev. Biol.* **14**, 26.

Jin, H. Stojnic, R., Adryan, B., Ozdemir, A., Stathopoulos, A. and Frasch, M. (2013). Genome-wide screens for in vivo Tinman binding sites identify cardiac enhancers with diverse functional architectures. *PLoS Genet.* **9**, e1003195.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E. H., Birney, E. and Furlong, E. E. M. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**, 473-486.

Kvon, E. Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J. O., Pagani, M., Schernhuber, K., Dickson, B. J. and Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* **512**, 91-97.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P. et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813-1831.

Li, Q., Brown, J. B., Huang, H. and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752-1779.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H. et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142-D147.

McKay, D. J. and Lieb, J. D. (2013). A common set of DNA regulatory elements shapes Drosophila appendages. *Dev. Cell* **27**, 306-318.

Narlikar, L. Sakabe, N. J., Blanski, A. A., Arimura, F. E., Westlund, J. M., Nobrega, M. A. and Ovcharenko, I. (2008). Genome-wide discovery of human heart enhancers. *Genome Res.* **20**, 381-392.

Ocorr, K., Fink, M., Cammarato, A., Bernstein, S. and Bodmer, R. (2009). Semi-automated optical heartbeat analysis of small hearts. *J. Vis. Exp.* **31**, pii: 1435.

Olson, E. N. (2006). Gene regulatory networks in the evolution and development of the heart. *Science* **313**, 1922-1927.

Peter, I. S. and Davidson, E. H. (2011). A gene regulatory network controlling the embryonic specification of endoderm. *Nature* **474**, 635-639.

Pfeiffer, B. D., Jenett, A., Hammonds, A. S., Ngo, T.-T. B., Misra, S., Murphy, C., Scully, A., Carlson, J. W., Wan, K. H., Laverty, T. R. et al. (2008). Tools for neuroanatomy and neurogenetics in Drosophila. *Proc. Natl. Acad. Sci. USA* **105**, 9715-9720.

Reim, I. and Frasch, M. (2010). Genetic and genomic dissection of cardiogenesis in the drosophila model. *Pediatr. Cardiol.* **31**, 325-334.

Seyres, D., Darbo, E., Perrin, L., Herrmann, C. and González, A. (2015). LedPred: an R/bioconductor package to predict regulatory sequences using support vector machines. *Bioinformatics* **32**, 1091-1093.

Steiner, F. A., Talbert, P. B., Kasinathan, S., Deal, R. B. and Henikoff, S. (2012). Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Res.* **22**, 766-777.

Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M. and Van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.* **3**, 1578-1588.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W. et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.

DEVELOPMENT